Statistical Science
1998, Vol. 13, No. 2, 136–141

Decision Analysis and Bioequivalence Trials

Dennis V. Lindley

Abstract. It is argued that the determination of bioequivalence involves a decision, and is not purely a problem of inference. A coherent method of decision-making is examined in detail for a simple trial of bioequivalence. The result is shown to differ seriously from the inferential method, using significance tests, ordinarily used. The reason for the difference is explored. It is shown how the decision-analytic method can be used in more complicated and realistic trials and the case for its general use presented.

Key words and phrases: Bioequivalence, utility, loss, expected loss, sample size, coherence, conjugacy.

A recent paper in this journal (Berger and Hsu, 1996, and its ensuing discussion, hereinafter referred to as BH) was on the topic of bioequivalence trials. The discussants included statisticians and practitioners in the pharmaceutical industry. There was nowhere any mention of an alternative approach to bioequivalence that is both simpler and more practically relevant than those in BH. The purpose of this note is to outline the principle behind the alternative, illustrate it on a simple example and indicate how it might be used in practice.

The basic idea in the alternative analysis is the recognition that the fundamental purpose of bioequivalence studies is the reaching of a decision. The final outcome is action; the generic product is, or is not, accepted as bioequivalent to the brand-name product. This is the simplest case. Complications can arise. For example, when the trial is inconclusive a decision may be made to extend it. This possibility was not discussed in BH and will not be included here, though the methods proposed extend to include this case. For clarity in presentation, we will consider only the simplest case of two decisions: to accept, or to reject, the claim of bioequivalence. Accepting the practical reality of the problem as one of decision, it is not obvious that the best way to reach a decision is through either significance tests or confidence intervals. Rather than attempt to find an optimum test, as in BH, it is surely more relevant to the work of the pharmaceutical industry and of regulatory agencies, to find an optimum decision procedure. Starting from this fundamental concept, it may turn out that significance tests are the best ways to proceed. The decision-based study suggests that this is not so, because test procedures use an unrecognized and possibly inappropriate assumption about the loss structure.

Bioequivalence trials concern decision-making under uncertainty. In the simplest case considered in BH, and the one to be analyzed here, the uncertain quantity upon which the decision depends is $\eta_T - \eta_R$, the difference between a mean η_T for the generic (test) drug and the corresponding mean η_R for the brand-name (reference) drug. This difference will be denoted by $\theta = \eta_T - \eta_R$. As explained in BH, if $\theta_L < \theta < \theta_U$, the drugs are sufficiently close to be declared bioequivalent. Here θ_L and θ_U are numbers laid down by the regulatory agency. The usual scenario has $\theta_L = -\theta_U$ with θ_U, now written as Δ, positive; $\theta = 0$ indicates perfect agreement in means between the two drugs, and Δ is the upper limit on any permitted disagreement. It will be convenient in what follows to equate bioequivalence to $\theta^2 < \Delta^2$, rather than $-\Delta < \theta < \Delta$. If θ were known, decision would be easy. The fact that it is not is a major cause of trouble.

There are two decisions, d_1 to declare bioequivalence, d_0 to deny it; and one uncertain quantity θ. A key consideration in any decision analysis is how the quality of a decision depends on the value of the uncertain quantity. As a partial statement of this, the regulatory agency thinks d_1 is better than d_0 if $\theta^2 < \Delta^2$, whereas the roles are reversed if $\theta^2 > \Delta^2$. This statement, on its own, is not sufficiently precise because it does not say how much better one decision is than another. Presumably if $\theta = 0$, d_1 is

Dennis V. Lindley was, before retirement, Head, Department of Statistics, University College London. He now lives at 2 Periton Lane, Minehead, Somerset, TA24 8AQ, England.

much better than d_0, but if θ^2 is only a little less than Δ^2, then the superiority of d_1 is only slight. The sensible way to describe the dependence of the quality of a decision on θ is through a utility function $u(d, \theta)$ which measures the worth, or utility, of d when the uncertain value is θ. Here there would be two utility functions, one for each of the decisions. Presumably $u(d_1, \theta) > u(d_0, \theta)$ for $\theta^2 < \Delta^2$, with the inequality on utilities reversed when $\theta^2 > \Delta^2$. The two functions reflect the regulatory agencies's view, but say much more in describing how much better one decision is than another. In a two-decision problem it is not necessary to specify both utility functions; it is enough to specify the difference between them for each value of the uncertain quantity. We shall work with $u(d_0, \theta) - u(d_1, \theta)$, the difference between the utility of rejection, and the utility of acceptance, of bioequivalence, and call it the loss of declaring bioequivalence (over that of denying it). It is written $L(\theta) = u(d_0, \theta) - u(d_1, \theta)$. Since only the one comparison is involved, that of d_0 above d_1, the notation $L(\theta)$ without reference to the decisions will suffice. Nevertheless it is important in what follows to remember that it is the loss involved in stating bioequivalence instead of denying it, rather than the other way round. In accord with the ideas of the agencies described above, $L(\theta)$ will be large if θ^2 is large, since it would be seriously wrong to conclude bioequivalence when η_T is very different from η_R, implying the magnitude of their difference θ is large. Equally, $L(\theta)$ is negative when $\theta^2 < \Delta^2$, since there is then a gain in asserting bioequivalence, and gain is negative loss. A possible form of loss function is shown by the continuous curve in Figure 1. It is symmetric about the origin, is continuous, increases with θ^2 and is bounded at infinities;

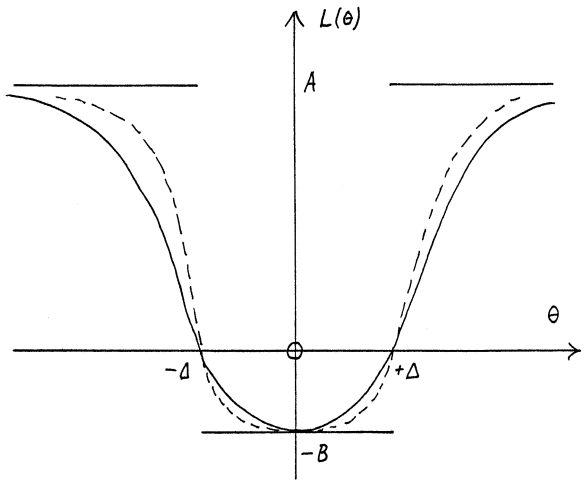

FIG. 1.

$L(-\Delta) = L(\Delta) = 0$ at the borderlines between the two decisions. Other types of loss function will be discussed later.

If bioequivalence is to be treated by decision analysis, then there is general agreement over the introduction of a loss function. This leaves the problem of handling the uncertainty of θ. If you knew the value of θ, the loss function would tell you unambiguously what to do. If you could confidently limit θ to a small interval, again you would have no problem unless the interval included $-\Delta$ or $+\Delta$. Generally, you need to combine your beliefs about θ, obtained as a result of the trial and other possible considerations, with your loss structure, to reach a decision. In BH beliefs are expressed in the form either of significance levels or confidence coefficients. The former concentrates on the greatest probability of declaring bioequivalence when it does not obtain, perhaps the most serious error. But what of the error of false denial of bioequivalence? And what of the dependence of these errors on θ? There is also the point, repeatedly demonstrated in the statistical literature, that significance tests can be seriously misleading as expressions of beliefs: see, for example, Berger and Delampady (1987) and Berger and Sellke (1987). These are not the Berger of BH.

The more sensible thing to do, both for theoretical and operational reasons, is to express your beliefs about θ in the way we usually express uncertainty, namely, in terms of probability—that is, to provide θ with a probability distribution, in the form of a density $p(\theta)$. This may be combined with the loss function by calculating the expected loss $\int L(\theta)p(\theta)\,d\theta$, and declaring for bioequivalence if, and only if, this is negative. Expected loss replaces the actual loss that would be used were θ known. This approach does not concentrate on one type of error, as does the significance test, but balances the errors and the bonuses in a simple, yet all-embracing, way. It is operationally appealing but additionally has the support of much research, starting with Ramsey, and continuing with De Finetti, Savage and many others. A recent reference is Bernardo and Smith (1994). An expository treatment is Lindley (1985). These writings show that the procedure just described, using expected loss, is the only sensible way of decision-making for a single decision-maker. In particular, for a given type of trial, whatever size of trial is selected, and whatever data is collected, the conclusions from the various trials will be consistent with one another. The technical term is "coherent." Berger and Delampady (1987), referred to above, show that this coherence is not achieved with some significance tests. Casella and Berger (1987) show that tests are not so misleading

when used with one-sided cases similar to those used in BH. We shall later in this paper suggest they are, though effectively coherent, based on a loss function that may be inappropriate. Let us see how the method works out in a simple case.

Consider a bioequivalence trial of the type described in Section 2.1 of BH. Here n observations are made which are, given θ, independent and identically normally distributed with mean θ. To simplify further from BH, it will unrealistically be supposed that the common variance is known. This simplifies the mathematics, enabling the final result to be exhibited clearly. It will later be indicated how more general cases may be handled. Under mild conditions, it follows that your beliefs about θ as a result of the trial can be expressed as a normal distribution with mean x and variance σ^2, say. The calculation of x and σ^2 will be considered later. It remains to describe the loss structure. The form in Figure 1 suggests, to a statistician, a normal density "upside down." The form

$$(1) \quad L(\theta) = A - (A + B)\exp(-(1/2)\theta^2/c^2)$$

with positive A and B has the right shape. For large θ, there is a loss A. For true bioequivalence, $\theta = 0$, there is negative loss $-B$. The critical values $\pm\Delta$ should exhibit zero loss, so

$$(2) \qquad \log[A/(A + B)] = -(1/2)\Delta^2/c^2.$$

The scale on which loss is measured is immaterial, so only the ratio of A to B matters. We take $A + B = 1$, when (2) shows that c is the only undefined quantity in the loss function (1), A being a function of it. The normal form fits well with the normal expression of beliefs. They are often said to be "conjugate."

The mathematics that follows, and in particular the final result (4), is expressed mainly in terms of c, for the simple reason that it is more easily understood in that form. From the practical viewpoint of the users, producer and regulatory agency, it is better to think about the ratio A/B of losses. Equation (2) expresses one quantity in terms of the other, enabling the practitioner to interpret results like (6) in terms that are more relevant to the basic problem, which is one of practical decision-making, not of mathematics. It remains to evaluate the expected loss

$$\int L(\theta)p(\theta)\,d\theta = A - (2\pi\sigma^2)^{-1/2} \cdot \int \exp\left[-(1/2)\left\{\theta^2/c^2 + (\theta - x)^2/\sigma^2\right\}\right] d\theta.$$

The usual completion of the square in θ for the expression in braces, and subsequent use of the normal integral, gives for the expected loss

$$(3) \quad A - c(c^2 + \sigma^2)^{-1/2}\exp[-x^2/2(c^2 + \sigma^2)].$$

As this is the expected loss in stating bioequivalence, the optimum decision is to declare bioequivalence if, and only if, (3) is negative. On rearrangement, and using (2) with $A + B = 1$, this gives

$$(4) \quad x^2 < (\sigma^2 + c^2)[\log\{c^2/(c^2 + \sigma^2)\} + \Delta^2/c^2].$$

That is, bioequivalence is only agreed when x is sufficiently close to zero, corresponding to perfect bioequivalence. The right-hand side of (4) describes precisely what is meant by "sufficiently close." One of the simpler methods in BH, TOST, when modified by the replacement of an estimated variance by a known one, declares bioequivalence when

$$(5) \qquad x^2 < (\Delta - z\sigma)^2,$$

where z is the upper 100α-percentile of a standard normal distribution. Here α is the preassigned maximum error probability of declaring the drugs to be bioequivalent, when in fact they are not. There is a slight difference in the interpretation of σ in the two approaches. In (5) it is a sampling standard deviation. In (4) it is a posterior standard deviation. In practice the two will only differ slightly, a point to be discussed later. Notice that the user of TOST only has to specify α (and Δ). The decision procedure requires c (and Δ). As (2) demonstrates, c may be replaced by the ratio A/B.

When the value of θ is known, corresponding to $\sigma = 0$, the two methods, (4) and (5), agree in asserting bioequivalence when $x^2 < \Delta^2$, as they must since our loss function was selected to be zero at Δ, the critical value used in TOST. Most trials will be reasonably extensive and will produce a small value of σ; that is, θ will be known with some precision. It is therefore of interest to compare (4) and (5) for small σ. Expanding the right-hand side of (4) in powers of σ yields

$$(6) \qquad x^2 < \Delta^2 + \sigma^2(\Delta^2/c^2 - 1) + o(\sigma^2),$$

whereas a similar treatment in (5) yields

$$(7) \qquad x^2 < \Delta^2 - 2z\Delta\sigma + o(\sigma).$$

The two methods agree in that as soon as some uncertainty about θ occurs, the interval of x in which bioequivalence is accepted reduces. (We shall see later that realistically $\Delta^2/c^2 < 1$.) They disagree in the amount of the reduction, since it depends on σ in TOST but only on σ^2 with the loss function. A numerical example is illuminating.

Without loss of generality, the scale of measurement may be chosen so that $\Delta = 1$. Years of statistical practice has put $\alpha = 0.05$, odds of 19 to 1. A possible choice for our loss function is to put $A = 0.95$, $B = 0.05$, thereby declaring that stating bioequivalence when the drugs are widely different is 19 times as serious as denying it when they are exactly equivalent. From (2) it follows that $c = 3.122$, substantiating the claim above that $\Delta^2/c^2 < 1$. If the trial yields a standard deviation σ of 0.2, one-tenth of the length of the interval within which bioequivalence is acceptable, then in the decision approach the interval of acceptance of bioequivalence reduces from $(-1, +1)$ when θ is known to $(-0.98, +0.98)$, which is hardly any change. The TOST interval is $(-0.66, +0.66)$, exhibiting a substantial effect of uncertainty. Why the difference between the two methods?

The question is easily answered for the case of small σ by recognizing that TOST is then effectively coherent, with loss function equal to A for $\theta^2 > \Delta^2$ and $-B$ for $\theta^2 < \Delta^2$. The expected loss is $A(1 - P) - BP$, where

$$P = \int_{-\Delta}^{+\Delta} [2\pi\sigma^2]^{-1/2} \exp\left[-\frac{(\theta - x)^2}{2\sigma^2}\right] d\theta$$
$$= \Phi\left[\frac{\Delta - x}{\sigma}\right] - \Phi\left[\frac{-\Delta - x}{\sigma}\right]$$

and Φ is the standard normal distribution function. Bioequivalence is declared whenever $P > A$. The upper limit of x for bioequivalence when σ is small only involves the first term in P; the second is negligible. Choosing $A = 0.95$ gives the upper limit to be $\Delta - z\sigma$, as (5) with $\alpha = 0.05$. The decision-maker's choice of A is numerically equivalent to the statistician's choice of α. The difference between (4) and (5) thus rests solely on the differences between the loss functions. Both are shown in Figure 1.

It is not the statistician's task to say what form the loss function should take. The statistician's role is to assist the decision-makers to articulate their value judgements in the form appropriate for the coherent procedure. It is the decision-maker's loss function, not the statistician's. In bioequivalence trials, the decision-makers are the producers (the pharmaceutical firm) and the consumers (presumably represented by the regulatory agency). The former will be concerned with the seriousness of rejecting a satisfactory drug, the latter with the consequences of accepting a bad one. In this light, let the two loss functions, the inverted normal and the two-level one of TOST, be compared. They are alike at the two extremes of very large θ and $\theta = 0$, or can be made alike by suitable choices of A and B. It is around the critical values $\theta = \pm\Delta$ that they differ substantially. The inverted normal used here is continuous in θ, whereas that implied by TOST has discontinuities at $\pm\Delta$, there abruptly moving between the two extreme values. It is this contrast between smooth and abrupt transitions between extremes that accounts for the differences of critical values. The following development provides an alternative demonstration of this result.

For a general loss function with normal uncertainty, the expected loss is

$$(2\pi\sigma^2)^{-1/2} \int \exp\left[-(1/2)(\theta - x)^2/\sigma^2\right] L(\theta)\, d\theta.$$

If $L(\theta)$ is expanded in a Taylor series about $\theta = x$, where the exponential has its maximum, the expected loss is seen to be

$$L(x) + (1/2)\sigma^2 L''(x) + o(\sigma^2) \tag{8}$$

so that the first two terms provide an approximation for small σ. The first term alone corresponds to the case, $\sigma = 0$, of known θ. It is therefore the second term that reflects the change due to small uncertainty about θ. This depends, apart from σ, on $L''(x)$. Consequently the second derivative of the loss function plays a critical role. In a sense, the discontinuous loss used in TOST has an infinite second derivative near $\theta = \pm\Delta$ as is seen by approximating it by a smooth curve. The inverted normal, by contrast, has a modest second derivative. In passing, note that use of the approximation (8) to the expected loss leads to the form (6) for the critical value of x, as direct calculation easily verifies. The occurrence of effectively an infinite second derivative in TOST means that (8) is not valid near $x = \pm\Delta$. In fact, the term in σ^2 is replaced by one in σ, equation (7).

A key question for the practitioner is therefore how large is the second derivative of the loss near $\pm\Delta$. The discontinuous form may be unreasonable because it says that, if θ is just a bit bigger than Δ, it is extremely serious to declare bioequivalence; whereas if it is just a bit smaller, it is serious to deny it. Such an abrupt change may not accord with practical requirements, and some smooth transition may be more acceptable. While the superiority of d_1 over d_0 is clear when $\theta = 0$, it is less clear when it is a little below Δ. My personal view is that the inverted normal may affect the transition too slowly and that a form corresponding to the broken curve in Figure 1 may be preferable. It represents an intermediate form between (6), which does not treat the uncertainty sufficiently seriously, and (7), which overreacts. But, as explained above, the choice rests with the practitioners. They should

recognize that whatever coherent procedure is used, some loss function is implied.

It is unreasonable to expect a practitioner in either the pharmaceutical industry or a regulatory agency to have to think about the second derivative occurring in approximation (8). What they might more reasonably do is to think about the behavior of the loss function in the neighborhood of $\theta = \Delta$ (and $-\Delta$). (Its behavior near $\theta = 0$ and for large values of θ^2 has already been discussed in terms of A and B.) For example, one might consider values of θ near Δ that are meaningfully different from Δ, say $\Delta + a$, $\Delta - b$ for positive a and b, and compare the losses at these two values. The former will be positive, the latter negative. Their ratio will indicate a reasonable slope (derivative) for the loss function near $\theta = \Delta$. For example, the dotted curve in Figure 1 has a larger ratio than the smooth, normal curve there. My personal guess is that the implicit choice in TOST of the largest possible value for this ratio is unrealistic. Careful consideration of losses at a few values like these should enable a smooth curve to be drawn and provide a loss function. The expected loss may then be evaluated numerically.

This paper has only considered a simple, special case of decision analysis applied to bioequivalence trials. We now turn to broader considerations. The methods used in BH involve the user specifying Δ, the critical value, and α, the significance level. Decision analysis requires, as a minimum, Δ, A and the second derivative of loss around $\pm\Delta$; A plays a similar role to α but instead of concentrating on just one error, balances the two. (Recall $A + B = 1$.) The second derivative is a new feature, not present in TOST or the other tests except as an implicit and very large value. However, it is possible to go much further and consider the loss function in more detail. For example, it may be that, when θ corresponds to a time difference, it is not serious if the test drug reaches the site quicker than the reference drug, but is serious if it is slower. This would suggest a loss function asymmetric about zero. Many possibilities can be encapsulated in the loss structure.

The description so far has concentrated on a single uncertain quantity θ but the method of analysis extends to two, or more, quantities θ, ϕ. It is necessary to specify a loss structure $L(\theta, \phi)$ which can incorporate how the two quantities may interact. This is usually accomplished by considering values of (θ, ϕ) that incur the same loss, giving contours of constant loss in the space of (θ, ϕ), and then attaching a loss to each contour in the manner of a single quantity. The analysis will then require a bivariate probability distribution for (θ, ϕ).

The density assumed in the simple univariate case above was $N(x, \sigma^2)$, derived from a random sample from a normal distribution. Here σ, the standard deviation, refers to θ, not to the data. It is not, as in TOST, the standard deviation of the sample mean, partly because it will incorporate knowledge prior to the trial, in addition to knowledge gained in the trial. The former may not be negligible. A firm is unlikely to propose a drug for testing unless it is reasonably confident of success. Equally the regulatory agency would expect sensible responsibility on the part of the firm. A suggestion would be for prior knowledge to correspond to a neutral position in which the drug was thought as likely to be bioequivalent as not. Accepting a prior normal distribution with mean $\theta = 0$, this would correspond to a prior standard deviation of $\tau_0 = \Delta/0.6745$ so that the probability that $\theta^2 < \Delta^2$ is $1/2$. If τ_1 is the sampling standard deviation of each of the n observations in the trial, the precisions add and $\sigma^{-2} = \tau_0^{-2} + n\tau_1^{-2}$. Similarly, if $\bar{x}$ is the sample mean, $x = (n\bar{x}/\tau_1^2)/(n/\tau_1^2 + \tau_0^{-2})$, a weighted average of $\bar{x}$ and zero, with weights equal to the precisions. For large n, the prior knowledge will have little effect but may be influential for modest values of n.

Although the principal thesis of this note is that bioequivalence trials lead to decision problems, and might be presented as such, the decision approach used here also incorporates a practically meaningful resolution, within the inference framework that BH and the discussants espouse. To illustrate, the result in the last paragraph says that, after the trial has been completed, the distribution of θ is normal with precision (the inverse of the variance) $\tau_0^{-2} + n\tau_1^{-2}$, denoted by σ^{-2}, and mean $\bar{x}(n\sigma^2/\tau_1^2)$, denoted by x. Consequently inferential statements of an easily understood form, avoiding the complexities of confidence intervals, can be made. For example, the probability that bioequivalence obtains is the area under this normal curve, mean x, variance σ^2, between $\pm\Delta$.

Even the simplest bioequivalence problem discussed here hides within it another decision problem, that of how large the trial should be. In our notation, the determination of the sample size n. How this should be resolved has recently been considered by Lindley (1997) and the reader is referred to that paper and its discussion. The major new consideration is to balance, just as did the two losses, here expressed through A and B, with the cost of experimentation. This paper is in an issue of *The Statistician* devoted to sample size determination. One idea that might be useful in bioequivalence studies is for the industry to choose n using its own prior, recognizing that it will need to con-

vince the agency, which will be using its own prior. In the related context of acceptance sampling, this feature has been studied by Lindley and Singpurwalla (1991).

May I invite all concerned with bioequivalence trials to contemplate seriously the use of decision analysis in place of significance tests. It provides a simpler, more flexible and more realistic treatment of the problem. It requires the interested parties getting together and settling on a loss function. This function should be openly declared so that the procedure is open to constructive criticism. It is also convenient to agree on the form of prior knowledge, which mainly consists in bringing together relevant information about the drugs. In default of agreement, robustness studies may be used. Third, the uncertain structure of the data will be needed, exactly as in the methods in BH. Finally, for the observed data set (the sign of) the expected loss has to be calculated. There are many excellent computer packages that will perform the necessary integration. Here is a method that will embrace any loss, any prior and any form of data.

REFERENCES

BERGER, J. O. and DELAMPADY, M. (1987). Testing precise hypotheses (with discussion). *Statist. Sci.* **2** 317–352.

BERGER, J. O. and SELLKE, T. (1987). Testing a point-null hypothesis: the irreconcilability of *P*-values and evidence (with discussion). *J. Amer. Statist. Assoc.* **82** 112–139.

BERGER, R. L. and HSU, J. C. (1996). Bioequivalence trials, intersection–union tests and equivalence confidence sets (with discussion). *Statist. Sci.* **11** 283–319.

BERNARDO, J. M. and SMITH, A. F. M. (1994). *Bayesian Theory*. Wiley, Chichester.

CASELLA, G. and BERGER, R. L. (1987). Reconciling Bayesian and frequentist evidence in the one-sided testing problem. *J. Amer. Statist. Assoc.* **82** 106–111.

LINDLEY, D. V. (1985). *Making Decisions*. Wiley, Chichester.

LINDLEY, D. V. (1997). The choice of sample size (with discussion). *The Statistician* **46** 129–166.

LINDLEY, D. V. and SINGPURWALLA, N. D. (1991). On the evidence needed to reach agreed action between adversaries, with application to acceptance sampling. *J. Amer. Statist. Assoc.* **86** 933–937.